機械学習を用いたゲノムワイド遺伝子多型情報に基づくうつ状態のリスク予測

（書式１２）

氏　　名	高橋（たかはし）　雄太（ゆうた）
学位の種類	博士（医学）
学位授与年月日	2020年3月25日
学位授与の条件	学位規則第4条第1項
研究科専攻	東北大学大学院医学系研究科（博士課程）　医科学専攻
学位論文題目	機械学習を用いたゲノムワイド遺伝子多型情報に基づくうつ状態のリスク予測
論文審査委員	主査　教授　富田　博秋 教授　栗山　進一　　教授　菅原　準一

論文内容要旨

【背景・目的】ゲノムワイド遺伝子多型データを用いてうつ状態のリスクを予測したこれまでの研究では、あまり高い予測精度は示されていない。その原因の一つとして、表現型に実際には効果を持たないバリアントが多く予測モデルに含まれることにより、予測モデルの学習の段階では見かけ上、高い予測精度が得られるのに、独立したテストデータで検証すると低い予測精度しかえられない、過剰適合と呼ばれる現象が起きてしまうことがある。STMGP（Smooth-Threshold Multivariate Genetic Prediction）法は過剰適合を軽減させることで予測精度を向上させることを目的に開発された、機械学習を用いた予測モデルである。今回の研究では、ゲノムワイド遺伝子多型データからのうつ状態の予測にSTMGP法を用いて、予測精度を従来法と比較した。

【方法】東北メディカル・メガバンクプロジェクトでリクルートされた、宮城県在住の3,685人の遺伝子多型情報を用いて予測モデルを学習させ、岩手県在住の3,048人の遺伝子多型情報を用いて、予測モデルの精度を評価した。HumanOmniExpressExome BeadChip Arrayを用いてゲノタイピングした。うつ症状はCenter for Epidemiologic Studies-Depression Scaleで評価した。STMGP法による予測精度と過剰適合の程度は、遺伝子スコア法、GBLUP（genomic best linear unbiased prediction）法、SBLUP（summary-data-based best linear unbiased prediction）法、BayesR法、Ridge回帰法と比較した。

【結果】STMGP法による予測精度（predictive correlation coefficients ± 標準誤差）は0.0769 ± 0.0173であり、遺伝子スコア法（0.0332 ± 0.0178）、GBLUP法（0.0309 ± 0.0178）、SBLUP法（0.0164 ± 0.0178）、BayesR法（0.0100 ± 0.0185）、Ridge回帰法（0.0260 ± 0.0178）よりも高かった。またSTMGP法ではトレーニングデータでの見かけ上の予測精度は0.3232 ± 0.0153であり、遺伝子スコア法（0.9027 ± 0.0076）、GBLUP法（0.9627 ± 0.0017）、SBLUP法（0.9554 ± 0.0019）、BayesR法（0.9633 ± 0.0015）、Ridge回帰法（0.9998 ± 0.0000）よりも低く、過剰適合を軽減させることで予測精度が向上していることが示唆された。

【結論】STMGP法は過剰適合を軽減することで、従来法よりもゲノムワイド遺伝子多型データからうつ症状を予測する際に、高い精度を示した。微小な効果をもつ遺伝子多型の集合で説明されるような複雑な遺伝疾患の脆弱性の予測にSTMGP法が有効であることが示唆された。

【所定様式⑬No.1】大学院の課程による者（課程博）

審　査　結　果　の　要　旨

博士論文題目　機械学習を用いたゲノムワイド遺伝子多型情報に基づくうつ状態のリスク予測

所属専攻・分野名　医科学　専攻　・　精神神経学　分野

学籍番号　B6 MD 5081　氏名　高橋　雄太

本博士論文の背景として、うつ状態と相関するゲノムワイド遺伝子多型解析は多くなされてきているが、そのデータを用いてうつ状態のリスクを予測する研究では、これまでのところあまり高い予測精度は示されていないことがある。その原因の一つとして、うつ状態という表現型に実際には効果を持たない遺伝子多型を多く予測モデルに含むことにより、予測モデルを形成する学習の段階では見かけ上、高い予測精度が得られるものの、独立したテストデータで検証すると低い予測精度しかえられない、過剰適合と呼ばれる現象が起きてしまうことがあげられる。本論文は、この状況をうけ、過剰適合を軽減させることで予測精度を向上させることを目的に開発されたSTMGP（Smooth-Threshold Multivariate Genetic Prediction）法をうつ状態予測の機械学習に導入し、ゲノムワイド遺伝子多型データからのうつ状態の予測精度を従来法と比較したものである。東北メディカル・メガバンクプロジェクトでリクルートされた、宮城県在住の3,685人のHuman OmniExpress Exome BeadChip Arrayによる遺伝子多型情報を用いて予測モデルを学習させ、岩手県在住の3,048人の遺伝子多型情報を用いて、Center for Epidemiologic Studies-Depression Scaleで評価したうつ状態の予測モデルの精度を評価した。このSTMGP法による予測精度と過剰適合の程度を、遺伝子スコア法、GBLUP（genomic best linear unbiased prediction）法、SBLUP（summary-data-based best linear unbiased prediction）法、BayesR法、Ridge回帰法と比較した。STMGP法による予測精度（predictive correlation coefficients ± 標準誤差）は0.0769 ± 0.0173であり、遺伝子スコア法（0.0332 ± 0.0178）、GBLUP法（0.0309 ± 0.0178）、SBLUP法（0.0164 ± 0.0178）、BayesR法（0.0100 ± 0.0185）、Ridge回帰法（0.0260 ± 0.0178）よりも高かった。またSTMGP法ではトレーニングデータでの見かけ上の予測精度は0.3232 ± 0.0153であり、遺伝子スコア法（0.9027 ± 0.0076）、GBLUP法（0.9627 ± 0.0017）、SBLUP法（0.9554 ± 0.0019）、BayesR法（0.9633 ± 0.0015）、Ridge回帰法（0.9998 ± 0.0000）よりも低く、過剰適合を軽減させることで予測精度が向上していることが示唆された。STMGP法は過剰適合を軽減することで、従来法よりもゲノムワイド遺伝子多型データからうつ症状を予測する際に、高い精度を示すことが証明された。本研究はゲノム情報に基づくうつ状態の予測精度の向上を行なったのみならず、微小な効果をもつ遺伝子多型の集合で説明されるような複雑な遺伝疾患の脆弱性の予測にSTMGP法が有効であることを示唆したことでも意義が認められる。よって、本論文は博士（医学）の学位論文として合格と認める。